Express Mail Label No. EL 542092752US
Docket No. SURR.78

MEDIAN FILTER FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA

RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 60/253,178, "Informatics System," filed 11/27/2000, and U.S. Provisional Application No. 60/314,996, "Nonlinear Filter For Liquid Chromatography-Mass Spectrometry Data," filed 8/24/2001, both of which are herein incorporated by reference.

FIELD OF THE INVENTION

[0002] This invention relates generally to analysis of data collected by analytical techniques such as chromatography and spectrometry. More particularly, it relates to a nonlinear filter such as a moving median filter for noise reduction in mass chromatograms acquired by liquid chromatography-mass spectrometry.

BACKGROUND OF THE INVENTION

[0003] Liquid chromatography-mass spectrometry (LC-MS) is a well-known combined analytical technique for separation and identification of chemical mixtures. Chromatography separates the mixture into its constituent components, and mass spectrometry further analyzes the separated components for identification purposes.

[0004] In its basic form, chromatography involves passing a mixture dissolved in a mobile phase over a stationary phase that interacts differently with different mixture constituents. Components that interact more strongly with the stationary phase move more slowly and therefore exit the stationary phase at a later time than components that interact more strongly with the mobile phase, providing for component separation. A detector records a property of the exiting species to yield a time-dependent plot of the property, e.g., mass or concentration, allowing for quantification and, in some cases, identification of the species. For example, an ultraviolet (UV) detector measures the UV

absorbance of the exiting analytes over time. When liquid chromatography is coupled to mass spectrometry, mass spectra of the eluting components are obtained at regular time intervals for use in identifying the mixture components. Mass spectra plot the abundance of ions of varying mass-to-charge ratio produced by ionizing and/or fragmenting the eluted components. The spectra can be compared with existing spectral libraries or otherwise analyzed to determine the chemical structure of the component or components. Note that LC-MS data are two-dimensional; that is, a discrete data point (intensity) is obtained for varying values of two independent variables, retention time and mass-to-charge ratio (m/z).

[0005] LC-MS data are typically reported by the instrument as a total ion current (TIC) chromatogram, the sum of all detected ions at each scan time. Peaks in the chromatogram represent separated components of the mixture eluting at different retention times. A noise-free chromatogram **10,** shown in FIG. 1A, appears as a series of smooth peaks **12a-12c,** each extending over multiple scan times. As shown in the TIC chromatogram of FIG. 1B, however, LC-MS data often have high-intensity noise spikes **14a-14d** superimposed on the peaks. Although components elute over multiple scans, noise spikes typically do not extend beyond one scan time. If the TIC chromatogram has little noise, an operator can determine the total number of peaks and then examine each peak's corresponding mass spectrum to identify the eluted species. However, as the amount of noise present increases, it becomes more difficult for the operator to distinguish the chromatographic peaks, particularly if the noise level is higher than the signal level. In such cases, the operator is left to examine each individual mass spectrum manually, select the mass-to-charge ratios corresponding to known or likely mixture components, and then assemble a reduced total ion current chromatogram from the selected masses only. Such a procedure is clearly very time consuming. Furthermore, when the mixture contains unknown analytes, the operator cannot confidently determine which mass spectral peaks are noise and which are actual peaks. Thus the only recourse the operator has is to adjust various instrument parameters and repeat the experiment with a different sample, hoping for less noise in the resulting chromatogram.

[0006] Because it enables the identification and quantification of hundreds to thousands of analytes in a single injection, LC-MS is currently being used to analyze complex biological mixtures (see, e.g, D.H. Chace et al., "Mass Spectrometry in the Clinical Laboratory," Chem. Rev. 101 (2001): 445-477). Proteomics is a relatively new field that aims to detect, identify, and quantify proteins to obtain biologically relevant information. Both proteomics and metabolomics (the detection, identification, and quantification of metabolites and other small molecules such as lipids and carbohydrates) may facilitate disease mechanism elucidation, early detection of disease, and evaluation of treatment. Recent advances in mass spectrometry have made it an excellent tool for structural determination of proteins, peptides, and other biological molecules. However, proteomics and small molecule studies typically have a set of requirements that cannot be met by manual interpretation of the LC-MS data.

[0007] First, these studies require high-throughput analysis of small volumes of biological fluid. Manual data interpretation creates a bottleneck in sample processing that severely limits the number of samples that can be analyzed in a given time period. Furthermore, while large available sample volumes allow an operator to adjust parameters by trial and error to obtain adequate chromatograms and spectra, biological samples are available in such small volumes that it is imperative to extract useful information from all of the available sample. Second, unlike traditional research applications, in which a relatively small amount of data is required, the paradigm of these studies is to acquire enormous amounts of data and then mine the data for new correlations and patterns. Manual data analysis is therefore unfeasible. In addition, biological samples are generally complex mixtures of unknown compounds, and so it is not desirable to extract only known spectra and discard the remaining data, an approach that has been used for studies involving quantification of known compounds in a mixture. Finally, LC-MS instruments produce an enormous amount of data: a single one-hour chromatographic run can produce up to 80 MB of binary data. For storage and subsequent data mining purposes, it is highly desirable to reduce the amount of data to retain information while discarding noise. To satisfy these requirements, a data analysis method is needed that can acquire a large amount of data from low-volume biological mixtures, extract useful information from the

resulting noisy data set, and identify unknown compounds from the extracted information. An essential component of such a method is the ability to filter noise accurately so that peaks can be distinguished automatically.

[0008] The problem of filtering chromatographic noise has been addressed to various degrees in the prior art. The component detection algorithm (CODA) is an automated method for selecting mass chromatograms with low noise and low background. CODA is described in W. Windig et al., "A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry," *Anal. Chem.*, 68 (1996): 3602-3606. The method computes a smoothed and mean-subtracted version of each mass chromatogram, compares it with the original chromatogram, and calculates a similarity index between the two. Chromatograms whose similarity index exceeds a threshold value are retained and combined to form a reduced total ion chromatogram, while other chromatograms are rejected. CODA has proven very effective at selecting high-quality mass chromatograms. However, it can only accept or reject entire chromatograms based on their noise level, but cannot filter noise from an individual chromatogram. As a result, noisy chromatograms that contain useful information are eliminated, and important peaks may not be detected.

[0009] Techniques exist for filtering noise and background from spectrometric data. For example, U.S. Patent No. 5,995,989, issued to Gedcke et al., describes a filtering method in which an average background level and an average deviation from the background are computed and used to define a local threshold for each data point. Points exceeding the threshold are retained, while points below the threshold are considered to be noise and discarded. The technique described in Gedcke et al. is only effective for noise levels that are substantially below the level of the peaks. For data such as that illustrated in FIG. 1B, high-intensity noise spikes cannot be removed using the disclosed method.

[0010] In U.S. Patent No. 6,112,161, issued to Dryden et al., a method for enhanced integration of chromatography or spectrometry signals is described. A baseline signal is computed from a moving average of the actual signal. The difference between

the baseline and actual signal is a baseline-adjusted signal containing peaks and high-frequency noise. An intensity range of the noise is determined, and all signal outside of this range is considered to be peaks, while signal inside this range is considered to be noise. As with the method of Gedcke et al., the method of Dryden et al. can only be used when the noise intensity is substantially lower than the signal intensity. Because LC-MS data often has noise values exceeding the signal values, the method of Dryden et al. is not effective at removing noise from LC-MS data.

[0011] A moving median digital filter has been used to remove noise from mass spectrometry and potentiometric titration data, as described in C.L. do Lago et al., "Applying moving median digital filter to mass spectrometry and potentiometric titration," *Anal. Chim. Acta*, 310 (1995): 281-288. Each data point is replaced by the median of the values in a window surrounding the point. With respect to the mass spectrometry data, the filter is applied both to the electron multiplier output, i.e., the ion abundance values, and to the magnetic field sensor, i.e., the mass-to-charge ratio. The method is not, however, applied to two-dimensional data such as LC-MS data. In most cases, state-of-the-art LC-MS instruments do not report the mass spectra as continuous smooth peaks, but rather as centroided data, i.e., single-mass peaks at the average mass value of the true peak. Without centroiding, an unmanageable amount of data would be generated for each spectrum. A moving median filter applied to centroided mass spectral data would remove peaks and noise equally. Because the peak shape is removed in the reported data, filtering or analytical methods cannot be applied to the mass spectra. Moreover, in some cases, one major source of noise, detector noise, can corrupt an entire mass spectrum. If a high fraction of the points in the filter window are corrupted, then a median filter applied to the spectrum cannot remove this noise.

[0012] There is still a need, therefore, for a method for removing noise, particularly high-intensity spikes, from chromatographic and spectrometric data such as LC-MS data.

SUMMARY OF THE INVENTION

[0013] The present invention provides a method for filtering noisy mass chromatograms in two-dimensional liquid chromatography-mass spectrometry (LC-MS) data. The method can remove noise spikes that have a higher intensity than peaks corresponding to eluted analytes, while essentially retaining the peak intensity and shape. It therefore performs substantially better than conventional linear filters that assume a normal distribution of the noise.

[0014] In a method of the invention for characterizing a chemical or biological sample, a series of mass spectra are generated by chromatography (e.g., liquid chromatography) and mass spectrometry. A total ion current (TIC) chromatogram is obtained from the mass spectra, and a median filter is applied to the chromatogram, resulting in a filtered total ion chromatogram with lower noise than the original chromatogram. Preferably, individual chromatograms are generated from the series of mass spectra, and the filter is applied separately to each individual chromatogram. The total ion current chromatogram is then reconstructed from the individual filtered chromatograms. Alternatively, the raw mass spectral data can be compared with the filtered chromatograms and the raw data replaced with the corresponding filtered data if the filtered data has a lower intensity value. The TIC chromatogram can then be assembled from the thresholded raw data. Subsequent to filtering, additional post-acquisition processing steps can be performed, such as applying a component detection algorithm to the filtered data to select relatively noise-free individual chromatograms.

[0015] The filter can be any suitable median filter such as a moving median filter or modified median filter, and the method preferably also includes selecting and optimizing one or more parameters of the filter. For example, the parameter can be selected based on the scan rate of the mass spectrometer or on subsequent data analysis, such as peak selection, of the mass spectra.

[0016] Also provided by the present invention is a program storage device accessible by a processor and tangibly embodying a program of instructions executable by the processor to perform method steps for the above method.

BRIEF DESCRIPTION OF THE FIGURES

[0017] FIGS. 1A-1B are schematic diagrams of a noise-free total ion current chromatogram and a noisy total ion current chromatogram, respectively, as known in the prior art.

[0018] FIGS. 2A-2B are histograms of LC-MS spectral intensities of a five-protein mixture and a high-molecular weight human serum fraction, respectively.

[0019] FIGS. 3A-3B illustrate a moving median filter of the invention applied to a chromatogram peak and to a chromatogram noise spike, respectively.

[0020] FIG. 3C illustrates a moving average filter applied to a chromatogram noise spike.

[0021] FIG. 4 is a flow diagram of a median filter method of the present invention.

[0022] FIG. 5 illustrates the application of a moving median filter with a poorly chosen window size to a chromatogram peak.

[0023] FIGS. 6A-6B show total ion chromatograms, base peak traces, and two-dimensional LC-MS data obtained from an LC-MS experiment of a proteolytic digest of human serum, before and after application of the moving median filter of the invention, respectively.

[0024] FIG. 6C shows the total ion chromatogram, base peak trace, and two-dimensional LC-MS plot of FIG. 6A after application of a mean filter.

[0025] FIG. 7 is a block diagram of a hardware system for implementing the method of FIG. 4.

DETAILED DESCRIPTION OF THE INVENTION

[0026] The present invention provides a method for filtering chromatographic and spectrometric data to reduce noise in individual mass chromatograms, thereby facilitating subsequent selection of peaks or high quality chromatograms for component detection. In liquid chromatography-mass spectrometry (LC-MS) or gas chromatography-mass spectrometry (GC-MS) data, the noise intensity is often larger than the intensity of the

peaks corresponding to eluted species, making it very difficult to extract meaningful information from the data. The method of the invention is able to reduce substantially such large magnitude noise spikes.

[0027] LC-MS noise originates from a variety of sources corresponding to different components of the system. Each physical mechanism of the system contributes its own noise distribution to the final measured ion current. For example, chemical noise results from column bleed, i.e., long-time elution of strongly-adsorbed species at particular mass-to-charge ratios, low-concentration sample contaminants, and detection of the chromatographic mobile phase. In the mass spectrometer, the ion generation, selection, and detection processes all generate noise. Electronic signal processing and analog-to-digital conversion add noise to the acquired data. The noise sources and distributions are not well understood for all components, making it difficult to select an appropriate filter.

[0028] In order to understand the noise distribution in LC-MS data better, the present inventor has prepared histograms of non-normalized intensity distributions of LC-MS spectra of two samples. FIG. 2A is a histogram of log intensity for a tryptic digest of a five-protein mixture (equal amounts by mass of horse myoglobin, bovine RNAse A, bovine serum albumin, bovine cytochrome C, and human hemoglobin). FIG. 2B is a histogram of log intensity for a tryptic digest of a high-molecular weight fraction of human serum produced by ultrafiltration through a 10 kD cutoff membrane. In both cases, the samples were reduced with dithiothreitol and carboxymethylated using iodoacetic acid and sodium hydroxide prior to digestion. The two samples display similar distributions, one at relatively low intensity and the other at higher intensity. In general, the major noise component of LC-MS data is chemical. It is believed that the second distribution, at lower ion current intensities, is associated with electrical noise. Both distributions appear close to normal in the log scale.

[0029] In methods of the invention, a digital median filter, a nonlinear filter, is applied to individual mass chromatograms to reduce the noise level. A mass

chromatogram is a plot of intensity versus retention time for a particular range of mass-to-charge ratio of detected ions. A nonlinear filter applies a nonlinear function to the data to be filtered. A nonlinear filter such as a median filter is particularly well-suited to LC-MS data because of the noise distribution characteristics of this type of data. As recognized by the present inventor, noise in LC-MS data is not normally distributed in the linear scale, i.e., does not follow a Gaussian distribution. Additionally, empirical study of LC-MS data has revealed that in individual mass chromatograms, noise spikes typically occur over a single scan time only. Standard filtering techniques for LC-MS data use moving average filters, which are linear filters and therefore effective at removing only normally distributed noise. Note that while noise may be correlated in adjacent points along the mass axis of a mass spectrum, it is typically not as highly correlated along the time axis. Thus there is no guarantee that noise reduction methods developed for one of the dimensions can be extended to the other dimension.

[0030] A simple median filter used in the preferred embodiment of the invention is a moving median filter, illustrated in FIGS. 3A-3B. A moving median filter replaces each point with the median of the points in a window of a given size centered on the selected point. For example, a three-point window examines a selected point and the neighboring point on each side of the selected point. Moving median filters are used for noise suppression in image processing but, to the knowledge of the present inventor, have not previously been applied to chromatographic data. FIG. 3A illustrates the application of a three-point moving median filter to a smooth chromatographic peak extending over multiple MS scan times. The top plot is the raw data, and the bottom plot is the filtered data. Points on the peak side slopes are necessarily the median of the three values in the window, and do not change upon application of the filter. The highest point of the peak is replaced by the larger of the two neighboring values. Thus the moving median filter flattens the peak slightly.

[0031] FIG. 3B illustrates the effect of the same three-point moving median filter on a single-point noise spike. The points surrounding the spike change little, if at all, but the noise spike is replaced by the higher of its two adjacent points, i.e., is completely

removed. FIGS. 3A-3B highlight the benefits of a moving median filter: it removes high-intensity noise spikes while retaining the sharpness of peak edges. In general, the filter removes features narrower than its half-width while retaining features wider than its width. A three-point moving average filter applied to the same noise spike is illustrated in FIG. 3C. In this case, each point is replaced by the mean of itself and its two surrounding points. The points at the edge of the noise spike are increased in value, while the spike itself is reduced significantly, but is still present. If the noise is of larger magnitude than the actual peaks, then the filtered noise is comparable to the peaks, and the filter is not effective in reducing noise.

[0032] A flow diagram of a method **20** of the invention for reducing noise in LC-MS or GC-MS data is shown in FIG. 4. First, in step **22,** the time-dependent mass spectra are acquired. Next, in step **24,** a mass chromatogram is generated for each integer mass in the entire set of mass spectra. For example, peaks at masses of 1321.7 and 1322.1 are summed and combined into the mass chromatogram for an integer mass of 1322. Alternatively, data points can be combined into mass ranges that do not necessarily correspond to integer values. In step **26,** the median filter is applied to each mass chromatogram generated in step **24.** Next, in an optional step **28,** a component selection algorithm such as CODA is applied to the filtered mass chromatograms.

[0033] Finally, the filtered mass chromatograms are combined into a reduced or filtered total ion current chromatogram in step **30.** One method is simply to sum the intensities at each time point. Alternatively, the raw mass spectral data obtained in step **22** can be thresholded using the filtered chromatograms. To do this, each raw data point is compared with its corresponding point in the filtered chromatograms. Recall that the raw data contain points at non-integer values of mass-to-charge ratio, while the filtered chromatograms contain points corresponding to ranges of mass values. If the intensity value of the raw data exceeds the value of the corresponding filtered point, then the original point is replaced by the filtered value. If not, it is retained.

[0034] The method **20** is typically implemented as part of an automated data analysis method for two-dimensional LC-MS or GC-MS. Data filtered according to the present invention may be subjected to, for example, peak recognition algorithms and structural identification algorithms. It is anticipated that the filtered data can be much more successfully analyzed by subsequent algorithms than can unfiltered data. In fact, one of the problems with the CODA method is that it removes noisy chromatograms, thereby also removing any information contained within the chromatograms.

[0035] Although the method **20** is best implemented by applying the median filter to the individual mass chromatograms and then combining the filtered chromatograms into a reduced total ion current chromatogram, the filter alternatively can be applied directly to the original total ion current chromatogram. In individual mass chromatograms, noise spikes typically occur over a single scan time only and are therefore effectively filtered using a moving median filter of the invention. In the total ion current chromatogram, however, spikes that occur at different masses but adjacent retention times can effectively merge to extend over multiple scan times and therefore pass the filter.

[0036] The median filters used in step **26** have parameters that are adjusted to achieve optimal filtering of the signal. The moving median filter, for example, has one parameter, the window size, the number of points over which the median is computed. The optimal window size is determined by a number of factors including the typical peak width and the scan rate. The peak width of LC-MS or GC-MS data varies with column conditions, flow rate, and mobile and stationary phases, among other factors. The window size should not be wider than the typical peak width, or peaks will be significantly distorted. FIG. 5 illustrates the use of a moving median window that is larger than the peak width. As shown, the peak base is approximately five points wide, while the filter window is nine points wide. The peak is essentially removed, and so this filter width is unacceptable. The filter width must be decreased to three points before the peak can survive the filter substantially unchanged.

[0037] In addition, the scan rate determines the density of points in the chromatogram and therefore also affects the optimal window size. If scans are performed half as frequently as in the chromatogram of FIG. 3A, all else being equal, the peak contains fewer points and therefore a smaller window is needed to retain the peak while eliminating noise spikes. From the point of view of the present invention, therefore, it is desirable to scan more frequently, assuming the increased sampling does not result in degraded signal-to-noise ratios in the individual scans. In a preferred embodiment, the method derives an expected peak width from the chromatography parameters and resolution and then selects a window size based on the expected peak width.

[0038] Additionally, the optimal parameters are determined by the quality of the resulting reduced total ion chromatogram and the ease and accuracy with which subsequent component detection or automated peak picking can be performed.

[0039] In some embodiments, an adaptive window size is employed. The window size is not the same for all data points, but varies based on a number of factors. The window size can be selected based on characteristics of the data by analyzing subsets of points in each mass chromatogram. Alternatively, the variation in window size can be predetermined based on knowledge of the instrument conditions. If peaks at later retention times are known to be broader than peaks at earlier times, then the window is preset to increase with retention time.

[0040] Additional median filters include a modified median filter. This filter has more parameters than the moving median filter, but the parameters are optimized based on the same principles.

[0041] Methods of the invention preferably use standard algorithms for implementing the various steps. For example, the moving median filter is applied using existing techniques for obtaining the median of a set of points. In one such algorithm, the median window is applied sequentially to the data beginning at the lowest-time data point. Points within the window are ordered and the central point selected as the median. For

subsequent points, the earliest-time point is removed and the new point inserted into the correct position in the ordered set. At the edges of the data set, additional points are appended so that the window can be centered on the first and last points. Preferably, the additional points have the same values as the edge points.

[0042] An example application of the method is shown in FIGS. 6A and 6B. FIG. 6A shows a total ion current chromatogram, base peak trace, and two-dimensional plot acquired from an LC-MS experiment of a proteolytic digest of human serum. The darkness of each point in the two-dimensional plot corresponds to the detected intensity at that mass-to-charge ratio and retention time. Each point in the TIC is the sum of all points directly below it in the two-dimensional plot, while each point in the base peak trace is the maximum value of all points below it. A moving median filter with a window size of seven points was applied to the mass chromatograms extracted from the data. FIG. 6B shows the resulting filtered data. FIG. 6C shows the results of applying a seven-point mean filter to the same data. Note that the mean filter changes the data very little, while the median filter clearly brings out six smooth peaks.

[0043] Although not limited to any particular hardware configuration, the present invention is typically implemented in software by a system **40,** shown in FIG. 7, containing a computer **42** in communication with an analytical instrument, in this case a LC-MS instrument **44** that includes a liquid chromatography instrument **46** connected to a mass spectrometer **48** by an interface **50.** The computer **42** acquires raw data directly from the instrument **44** via an analog-to-digital converter. Alternatively, the invention can be implemented by a computer in communication with an instrument computer that obtains the raw data. Of course, specific implementation details depend on the format of data supplied by the instrument computer. Preferably, the entire process is automated: the user sets the instrument parameters and injects a sample, the two-dimensional data are acquired, and the data are filtered for subsequent processing or transfer to a suitable database.

[0044] The computer **42** implementing the invention typically contains a processor **52,** memory **54,** data storage device **56,** display **58,** and input device **60.** Methods of the invention are executed by the processor **52** under the direction of computer program code stored in the computer **42.** Using techniques well known in the computer arts, such code is tangibly embodied within a computer program storage device accessible by the processor **52,** e.g., within system memory **54** or on a computer readable storage medium **56** such as a hard disk or CD-ROM. The methods may be implemented by any means known in the art. For example, any number of computer programming languages, such as Java, C++, or LISP may be used. Furthermore, various programming approaches such as procedural or object oriented may be employed.

[0045] It is to be understood that the steps described above are highly simplified versions of the actual processing performed by the computer **42,** and that methods containing additional steps or rearrangement of the steps described are within the scope of the present invention.

[0046] It should be noted that the foregoing description is only illustrative of the invention. Various alternatives and modifications can be devised by those skilled in the art without departing from the invention. Accordingly, the present invention is intended to embrace all such alternatives, modifications and variances which fall within the scope of the disclosed invention.